INSTITUTE OF PUBLIC HEALTH

COLLEGE OF MEDICINE AND HEALTH SCIENCES

UNIVERSITY OF GONDAR

APPLICATION OF DATA MINING TO EXPLORE THE PATTERN OF TUBERCULOSIS: THE CASE OF DEBIRE BIRHAN HOSPITAL, NORTH SHOA, ETHIOPIA.

BY

MENGISTU YILMA

ADVISORS

1. Mr. Takele Tadesse (BSC, MPH)
2. Dr. Million Meshesha (PHD)

A THESIS SUBMITTED TO INSTITUTE OF PUBLIC HEALTH, COLLEGE OF MEDICINE AND HEALTH SCIENCES, UNIVERSITY OF GONDAR IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF PUBLIC HEALTH.

JUNE, 2012

GONDAR, ETHIOPIA

INSTITUTE OF PUBLIC HEALTH

COLLEGE OF MEDICINE AND HEALTH SCIENCES

UNIVERSITY OF GONDAR

APPLICATION OF DATA MINING TO EXPLORE THE PATTERN OF TUBERCULOSIS: THE CASE OF DEBIRE BIRHAN HOSPITAL, NORTH SHOA, ETHIOPIA.

BY: MENGISTU YILMA

Approved by the examining board

_____

Head, School of public Health

ADVISORS

1. Mr. Takele Tadesse (BSC, MPH) _____
2. Dr. Million Meshesha (PHD) _____

_____ _____

Examiner

ACKNOWLEDGEMENT

First and foremost, I would like to give my special thank to my advisors Dr. Million Meshesha and Mr. Takele Tadesse who gives me their unconstrained advice and support for the completion of this thesis. I also want to express my special thanks to University of Gondar for giving me this opportunity.

I would like to give my special thank to Ato Sharew who is working in Tb room in Debirebirhan hospital and the hospital manager and other staffs who was gave me their cooperation to access the data. I thanks very much for all participated in this research by data entry and supervision.

I want to give my heartfelt gratitude to my brother Sleshi Yilma who played a great role for my success. My gratefulness also goes to my kind wife Sr. Meseret Amare for her encouragement, support and patience during my study time.

I am also very much indebted to acknowledge department of health informatics and public health for their cooperation in any aspect of this work.

Lastly but not least I thanks all my families, friends and class mates for their support. Thanks God to all. Amen!!!

Table of Contents **page**

ACKNOWLEDGEMENT ii

List of annexes v

List of figures v

List of tables v

ABSTRACT vii

1. INTRODUCTION 1

1.1. Statement of the problem 1

1.2. Literature review 2

1.2.1. Data mining and other statistical tools 2

1.2.2. Application of data mining in healthcare 3

1.3. Justification of the study 6

2. OBJECTIVES 7

2.1. General objective 7

2.2. Specific objectives 7

3. METHOD AND MATERIALS 8

3.1. Study design 8

3.2. Study area and period 9

3.3. Study population 10

3.4. Study variables 10

3.5. Sample size 11

3.6. Data collection procedures and data quality control 11

3.7. Data processing and management 12

4. ETHICAL CONSIDERATION 13

5. DISSEMINATION OF RESULTS 13

6. RESULTS 14
6.1. Characteristics of the registered patients 14
6.2. Model building experiments 16
6.3. Exploring pattern 26
6.4. Comparison of experimental results 28
7. DISCUSSION 29
7.1. The model 29
7.2. Exploring the pattern 30
7.3. Selection and evaluation of the model 31
8. STRENGTH AND LIMITATION OF THE STUDY 32
8.1. Strength of the study 32
8.2. Limitations of the study 32
9. CONCLUSION AND RECOMMENDATION 33
9.1. Conclusion 33
9.2. Recommendation 34
10. REFERENCES 35
Declaration 47

List of annexes

Annex 1_ Description of the dataset 37
Annex 2_Information Sheet and Consent Form 40
Annex 3- Sample J48 pruned tree 45

List of figures

Figure 1- CRISP-DM process model(5) 8
Figure 2- Map of study area 9
Figure 3- An exploratory analysis of age category of TB patients in Debirebirhan hospital 15
Figure 4- J48 decision tree algorithm on training 18
Figure 5- MLP algorithm under training 24

List of tables

Table 1. List of attributes, their possible values and description 10
Table 2: The characteristics of patients registered for treatment in Debirebrhan hospital, October 2001 to June 2011. 14
Table 3: Treatment outcome of patients registered for treatment in Debirebirhan hospital October 2001 to June 2011. 16
Table 4- Detailed Accuracy by Class of J48 decision tree 18
Table 5- Confusion Matrix of J48 decision tree 19
Table 6- Detailed Accuracy by Class for J48 decision tree 21
Table 7- Confusion Matrix for J48 decision tree with percentage split 21
Table 8- Detailed Accuracy by Class for SMO 23
Table 9- Confusion Matrix for SMO 23
Table 10-Detailed Accuracy by Class for MLP 25
Table 11- Confusion Matrix for MLP 26
Table 12- Comparison of experimented results 28

LIST OF ACRONYMS

CRISP-DM	Cross Industry Standard process for Data Mining
CSC	Cost sensitive classifier
DM	Data Mining
DOTS	Direct observed treatment short course
EBM	Evidence Based Medicine
GNU	General Public License
GRNN	General Regression Neural Network
HSDP	Health Sector Development Program
MLNN	Multilayer Neural Network
MLP	Multilayer Perceptron
NA	Not Applicable
NB	Navey Bayes
PTB	Pulmonary Tuberculosis
RF	Randomforest
ROC	Receiver Operator Characteristics
RPTB	Retroviral Pulmonary Tuberculosis
SMO	Sequential minimal optimization
SOMTE	Synthetic Oversampling Minority Technique
SVM	Support Vector Machine
WEKA	Waikato Environment for Knowledge Analysis

ABSTRACT

Background: Tuberculosis is the leading cause of mortality among infectious diseases worldwide. Evaluation of treatment outcome is used as a major indicator of program quality performed by the health institutes. Since data mining can be applied to explore interesting, useful and task oriented knowledge from huge amount of data, this study implemented data mining to explore the pattern of tuberculosis and to develop predictive model in relation to the treatment outcome.

Objective: To explore patterns from the tuberculosis data and develop predictive model using data mining technology.

Methods: An open source data mining tool WEKA software was used in this study. The study design was the standard procedure to data mining called Cross Industry Standard process for Data Mining (CRISP-DM). A total of 4780 patient records were taken for this study from the registration book of tuberculosis patients registered for treatment in Debirebirhan hospital from October, 2001 to June, 2011.

Result: From the total 4780 registered patients 1320 (27.6%) were perform HIV test and from those 468 (35.6%) were reactive for HIV. From pulmonary positive tuberculosis cases 668 (51.5%) patients were performed sputum follow up test at 7^{th} month. The outcomes were cured 649 (50%), completed 1813 (37.9%), died 370 (7.7%), failed 4 (0.3%), defaulted 458 (9.58%) and transferred out 1486 (31.1%). Multilayer perceptron registers the highest accuracy of 85.8%. All the attributes used in this study were considered as a predictor attributes to explore the pattern.

Conclusion and recommendation: All algorithms experimented in this study showed a promising result. Sputum test result of 7^{th} month for smear positive patients was the most determinant predictor attribute for cured and failed classes. Multilayer perceptron (MLP) was the best algorithm to classify and predict tuberculosis data. The outcomes died, defaulted and failed classes accounted 17.4% which is serious problem as a public health concern. Further research will be expected to be undertaken on large scale data and adding attributes like sign and symptom of the patients.

1. INTRODUCTION

1.1. Statement of the problem

Evaluation of treatment outcome is used as a major indicator of program quality performed by the health institutes(1). It enables the identification of problems, so that program managers, care providers and decision makers can institute appropriate action to overcome the problem and to improve the program performance(1). Treatment outcome can be the effect of any strength or weakness in relation of diagnostic, adherence or effectiveness of treatment. Tuberculosis is the leading cause of mortality among infectious diseases worldwide (2). If not treated properly, that can be fatal(3).

According to the 2011 WHO report, there were an estimated 8.8 million incident cases of TB, 1.1 million deaths among HIV-negative cases of TB and an additional 0.35 million deaths among people who were HIV-positive globally in 2010(2). Ethiopia is one of the 22 high tuberculosis burden areas. And the 7th high burden country with mortality of 35, prevalence of 394, and incidence of 261 estimated epidemiological burden of TB rate per 100000 populations in 2010(2). But the Ethiopian national tuberculosis survey 2010-2011 preliminary report showed prevalence of all forms of TB of 224/100,000 and incidence of sputum smear positive TB <70/100,000 population.

The target set for the prevention and control of TB have been to achieve 85% treatment success rate and a detection rate of 70% of new sputum positive TB cases. The national cure and treatment success rates are 67% and 84%, which is on track towards Health Sector Development Program III target, while the case detection rate remains at 34%, far less than what was planned for Health Sector Development Program III(4).

Since data mining can be applied to explore interesting, useful and task oriented knowledge from huge amount of data, this study implemented data mining to explore the pattern of tuberculosis and to develop predictive model in relation to the treatment outcome. The models developed using data mining technique can look the 10 years historical tuberculosis patients data feed to it, learn the pattern or

knowledge from that data, classify according to its class and predict what will happen in the future data. It is therefore the aim of this study to explore important pattern from tuberculosis data, build and test the predictive model by using this historical data.

To this end this study attempted to answer the following research questions:-

- What patterns in relation to the tuberculosis patients profile associated with the treatment outcome?
- How the effective predictive model can build and test using the existing historical data for the prediction of the future patterns?
- How can evaluate the developed model using different parameters to select the most accurate and effective model for our future prediction?

1.2. Literature review

An enormous proliferation of databases in almost every area of business has created a great demand for new, powerful tools for turning data into useful, task-oriented knowledge. In the efforts to satisfy this need, researchers have been exploring ideas and methods developed in machine learning, pattern recognition, statistical data analysis, data visualization, neural networks, etc. These efforts have led to the emergence of a new research area, frequently called data mining and knowledge discovery*(5).* One may think why use data mining today in healthcares when statistical analysis are already performed. The reasons are blasting of huge data (data is the raw material that fuels business growth—if only it can be mined), emerging of powerful computers and availability of facilities for data mining like free open source software(6).

1.2.1. Data mining and other statistical tools

The current Information age is characterized by an extraordinary growth of data that are being generated and stored about all kinds of human endeavors. An increasing proportion of these data is recorded in the form of computer databases, so that the computer technology may easily access it. The availability of very large volumes of such data has created a problem of how to extract useful, task-oriented knowledge.

Data analysis techniques that have been statistically used for such tasks include regression analysis, cluster analysis, numerical taxonomy, multidimensional analysis, other multivariate statistical methods, time series analysis, nonlinear estimation techniques, and others. These techniques have been widely used for solving many practical problems. They are, however, primarily oriented toward the extraction of quantitative and statistical data characteristics, and as such have inherent limitations(5).

We are today awash in data, primarily collected by governments and businesses. Automation produces an ever-growing flood of data, now feeding such a vast ocean that we can only watch the swelling tide, amazed. Surprised by our apparent inability to come to grasps with the knowledge swimming in the vast ocean before us, we know there must be a vast harvest to be had in this ocean, if only we could find the means that is data mining and knowledge discovery(7).

Data mining is a tool that assists business analysts with finding patterns and relationships in the data. Data mining does not replace statistical techniques. Rather, it is an extension of statistical methods that is in part the result of a major change in the statistics community(8).

One of the differences between Data Mining and statistics is that Data Mining automates the statistical process requiring in several tools. Statistical inference is assumption driven in the sense that a hypothesis is formed and tested against data. Data Mining, in contrast is discovery driven. That is, the hypothesis is automatically extracted from the given data. The other reason is Data Mining techniques tend to be more forceful for real-world messy data and also used less by expert users(9)

1.2.2. Application of data mining in healthcare

Data mining has been applied with success in different fields including marketing, banking, customer relationship management, engineering and various areas of science. However, its application to the analysis of medical data has been relatively limited. Thus, there is a growing pressure for intelligent data analysis such as data mining to facilitate the extraction of knowledge to support clinical specialists in making decisions(10). Medical datasets have reached enormous capacities. This data may contain valuable information that awaits extraction. The knowledge may be

encapsulated in various patterns and regularities that may be hidden in the data. Such knowledge may prove to be priceless in future medical decision-making(10)

The practice of using concrete data and evidence to support medical decisions (also known as evidence-based medicine or EBM) has existed for centuries, John snow use map to locate the site of the disease source, Florence Nightingale use polar-area diagrams to show that many army deaths could be traced to unsanitary clinical practices. But today, the size of the population, the amount of electronic data gathered, along with globalization and the speed of disease outbreaks make it almost impossible to accomplish what the pioneers did. This is where data mining become the most useful to health care.(11).

Data mining is the extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. It has ability to build a successful predictive model depends on past data. Data Mining is designed to learn from past success and failures and will be able to predict what will happen next (future prediction)(9)

Data classification process using knowledge obtained from known historical data has been one of the most intensively studied subjects in statistics, decision science and computer science. Data mining techniques have been applied to medical services in several areas, including prediction of effectiveness of surgical procedures, medical tests, medication, and the discovery of relationships among clinical and diagnosis data. In order to help the clinicians in diagnosing the type of disease computerized data mining and decision support tools are used which are able to help clinicians to process a huge amount of data available from solving previous cases and suggest the probable diagnosis based on the values of several important attributes. There have been numerous comparisons of the different classification and prediction methods, and the matter remains a research topic. No single method has been found to be superior over all others for all data sets(10).

One of the advantages of a classification model in the form of a decision tree is that it shows the relevant attributes selected by the algorithm (the attributes labeling the internal nodes of the tree) in a simple and intuitive hierarchical way, where attributes at the top of the tree are more relevant than attributes at the bottom. The attribute at

the root node, in particular, is considered the most relevant attribute for classification(12)

A study done in India to classify the tuberculosis patient's pattern by using support vector machine (SVM), C4.5 decision tree, NaiveBaye, K nearest neighbor (KNN), Bagging, AdaBoost, and RandomForest classifiers, the finding showed SVM had the highest accuracy (98.7%) followed by Bagging(98.4%) and RandomForet (98.3%) compared to other classifiers to classify the tuberculosis patients as pulmonary tuberculosis (PTB) and Retroviral Pulmonary Tuberculosis (RPTB) (10).

The study conducted on tuberculosis disease diagnosis using artificial neural network trained with genetic algorithm in Turkey, distinct parameters, which fall into three groups,: demographic variables, constitutional symptoms and radiographic findings, have been used for the study. The result showed Multilayer neural network (MLNN) with genetic algorithm had better performance (94.88%) for the diagnosis of tuberculosis compared to algorisms studied by other researchers(13).

The study done in India on Predictive models for anti-tubercular molecules using machine learning on high-throughput biological screening datasets using datasets for three confirmatory bioassay screens based on microplate Alamar blue assay for identifying inhibitors of Mycobacterium. These correspond to assay identifiers: AID1626, AID1949 and AID1332. All the three assays were conducted through the Tuberculosis Antimicrobial Acquisition and Coordinating Facility (TAACF)(14). The results found with different cost sensitive classifier (CSC) algorithms NB (Navey Bayes), RF (Random forest), and SMO (sequential minimal optimization) illustrated:

In dataset AID1332, AID1626 and AID1949 CSCRF showed best accuracy (82.57%), (82.36%) and (81.25%) respectively compare to other classifiers to develop a predictive model for anti-tubercular molecules in order to identify inhibitors of mycobacterium tuberculosis(14).

Another study was done in Nigeria to develop a decision support system for diagnosing TB using fuzzy method. The study evaluated the diagnosis of thirty patients using fuzzy methodology and the results gotten were in the range of the pre-defined limits by the domain experts. The developed decision support model can

diagnosed one patient with probability of 61% as sever tuberculosis compared to those of medical doctors(15).

The study done on treatment outcome of tuberculosis in Gondar university hospital in 2009, of 4000 tuberculosis patients who were registered at the hospital during the study period the result showed 1181(29.5%) had successfully treated, 730 (18.3%) defaulted, 403 (10.1%) died, 6(0.2%) treatment failure and 1680 (42.0%) transferred out.(16). Actually this study was not conducted using data mining technique but it is used in this study to compare the treatment outcome.

Since it is new and little is done for application of data mining in health care in Ethiopia, this study will aim to explore tuberculosis patterns based on treatment outcome and potential predictor attributes that can contribute for identifying important knowledge and encourage future researcher on this area.

1.3. Justification of the study

Data mining today is the set of procedures and techniques for discovering and describing patterns and trends in data. There is vast potential for data mining applications in the tuberculosis data. It is possible to explore the pattern from the existed data, build and test the effective model for prediction of the future pattern and evaluate the performance of the developed model using different parameters like Kappa measure, F-measure, ROC (Receiver Operator Characteristics) area, precision and recall.

This research initiated since there was no study conducted previously on application of data mining to explore the pattern of tuberculosis in Debirebirhan hospital.

The other motive to conduct this research is for the reason that data mining is powerful, applicable and feasible tool to automate and up-to-date the healthcares decisions. This can provide the health professional decision making support for easy detection, diagnosis and treatment. The planners and decision makers can also use to planning and resources allocation based on evidence.

2. OBJECTIVES

2.1. General objective

To explore patterns from the tuberculosis data and develop predictive model using data mining technology in Debirebirhan hospital

2.2. Specific objectives

- To identify patterns from the tuberculosis data.
- To develop a predictive model from tuberculosis data
- To identify an appropriate data mining task and algorithms.

3. METHOD AND MATERIALS

3.1. Study design

The study design was the standard procedure to data mining Cross Industry Standard process for Data Mining (CRISP-DM) which is the frame work guide provides a complete blue print to follow the data mining activity(5) was used. CRISP-DM has 6 phases as depicted in figure 1.

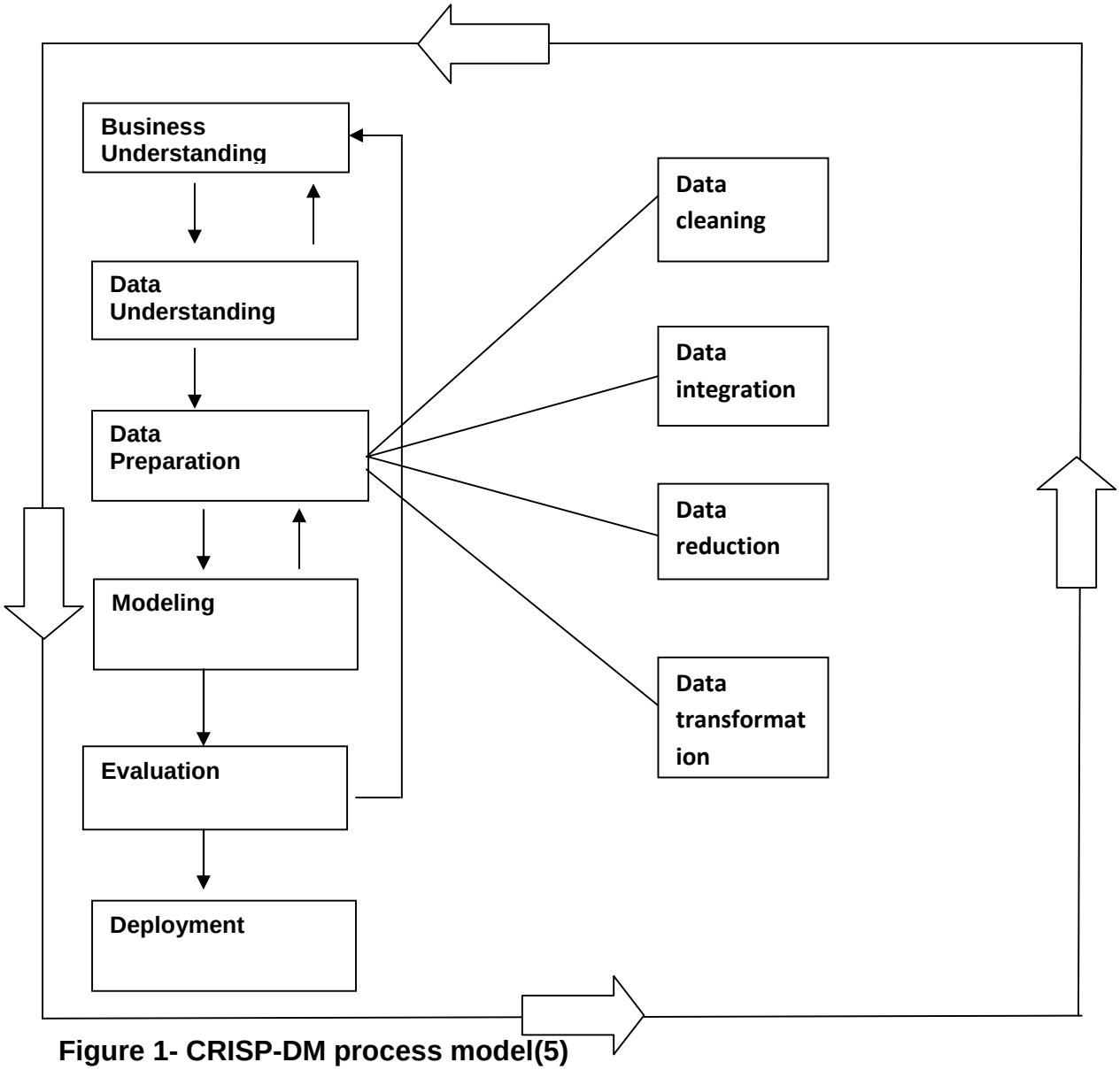

Figure 1- CRISP-DM process model(5)

3.2. Study area and period

Debirebirhan hospital is found in Amhara national regional state, North Shoa zone, Debirebirhan town, 718 km away from Bahir Dar and 130 km away from Addis Ababa. It was established in 1937 (1929 E.C) which was intended to serve for 20_25 thousand population but now it is serving almost for the zone population since it is the zonal referral hospital. There are 5 wards, 128 beds and more than 20 service giving rooms. North Shoa zone has 27 woredas with 52 town kebeles and 386 rural kebeles. The total populations of the zone in 2011/2012 are 2012224 (1016498 males and 995726 females).

This information enables us to assess and understand the situation related to the facts, assumptions, inputs, outputs and processes undertaken in the study. The study period was March 15, 2012 to May 15, 2012.

Map of study area

Figure 2- Map of study area

3.3. Study population

The whole patients of tuberculosis registered for treatment in Debirebirhan hospital starting from the beginning of DOTS (Directly Observed Treatment Short course) from October, 2001 to June, 2011 were the principal target dataset for this study.

3.4. Study variables

Table 1: List of attributes, their possible values and description

No	Attributes	Data type	Description	Values
1	AC	Categorical	Age category	1=0_14, 2=15_24, 3=25_34, 4=35_44, 5=45_54, 6=55_64, 7=65+
2	SEX	Categorical	Gender	1=male, 2=female
3	PTA	Categorical	Patient address	1=rural, 2=urban
4	TBC	Categorical	Tuberculosis case	1=p/pos Tb, 2=p/neg Tb, 3=Extra pulmonary Tb
5	WT	Numerical	Weight	Real
6	TRC	Categorical	Treatment category	1=new, 2=relapse, 3=treatment failure, 4=return after default, 5=transferred in, 5=others
7	TRI	Categorical	Treatment interruption	0=none, 1=less than 4 weeks, 2=4-8 weeks, 3=greater than 8 weeks
8	FST2M	Categorical	Follow up sputum test at 2^{nd} month	0=none, 1=yes, 2=not applicable
9	FST5M	Categorical	Follow up sputum test at 5^{th} month	0=none, 1=yes, 2=not applicable
10	FST7M	Categorical	Follow up sputum test at 7^{th} month	0=none, 1=yes, 2=not applicable
11	STR2M	Categorical	Sputum test result at 2^{nd} month	0=negative, 1=positive, 2=not applicable
12	STR5M	Categorical	Sputum test result	0=negative, 1=positive, 2=not

			at 5th month	applicable
13	STR7M	Categorical	Sputum test result at 7th month	0=negative, 1=positive, 2=not applicable
14	HIVTO	Categorical	HIV test offered	0=no, 1=yes
15	HIVTP	Categorical	HIV test performed	0=no, 1=yes, 3=not applicable
16	HIVTR	Categorical	HIV test result	0=nonreactive, 1=reactive, 3=indeterminate
17	CPT	Categorical	Cotrimoxazole prophylaxis treatment started	0=no, 1=yes,3=not applicable
18	EHIVC	Categorical	Enrolled in HIV care	0=no, 1=yes,3=not applicable
19	ART	Categorical	Anti retro viral treatment started	0=no, 1=yes,3=not applicable
20	TRO	Class	Treatment outcome	1=cured, 2=complete,3=died, 4=failed, 5=defaulted, 6=transferred out

3.5. Sample size

Since all the 10 years data starting from the beginning of DOTS program, calculating sample size was not needed.

3.6. Data collection procedures and data quality control

The data on registration book of the TB patients were entered in to the template prepared using Epi info. The task of data entry was carry out after giving training to 2 persons who have an IT (computer science diploma), 1 supervisors who have BSC in nursing and having good computer skill for supervising the quality of the data for two days. The training was focused on how much to be careful when entering the data in the template and the overall importance of the accuracy of the data for this research. The issue of confidentiality and secrecy was stressed during training. The objective of the study was explained to the trainees. The quality of data entry was

checked frequently by the supervisor and it was also followed attentively by the investigator closely. After the completion of entry the data transported from the Microsoft Access to excel file. After some adjustments the data in the excel file saved to CSV comma delimited then save to arff to make it appropriate for the WEKA software. Most of data cleaning work on missing and noisy data were done during data entry. The missing was filled empty in the template and filled with “?” in the excel table. The missing that was filled with “?” was automatically replaced using expected maximization (EM) imputation in the WEKA filter option.

3.7. Data processing and management

An open source data mining tool WEKA 3.7.5 (Waikato Environment for Knowledge Analysis) was used in this study. WEKA is a popular suite of machine learning and knowledge analysis software, developed at the University of Waikato.

To partition the dataset to training and testing dataset the 10 fold cross validation and percentage split was used for J48 decision tree. In the case of percentage split 65% of the data was used for training and the remaining 35% for testing(6)

For creating predictive modeling classification algorithms (6) such as J48 decision tree, a standard algorithm that is used widely for practical machine learning, SMO (sequential minimal optimization) algorithm for support vector machine which are an important new paradigm in machine learning, MLP (MultiLayerPerceptron) is an artificial neural network model that maps sets of input data onto a set of appropriate output was used in this study.

For evaluating the performance of the developed model WEKA software package has used different parameters. The kappa statistics tells us about the classification agreement of the actual value and the predicted value of the classes. When the kappa statistics is one shows the perfect agreement between the actual and the predicted values for the specified class. The value of the kappa statistic near to one means there are a better agreement. But when we evaluate the imbalanced data the accuracy and the kappa statistics can behave poorly to the minority class. If the data set is extremely imbalanced, even the classifier classifies all the majority instances

correctly and miss classify all the minority instances, the accuracy of the learner is still high because there are much more majority instances than minority instances. Under this circumstance accuracy can not reflect reliable prediction for minority class. In this condition ROC area is one of the popular metrics to evaluate the learners for imbalanced data set(17)

4. ETHICAL CONSIDERATION

The ethical clearance was gained from the Ethical review board of University of Gondar. Communication with the concerned bodies was made through that formal letter gained from the University of Gondar. After the purpose and objective of the study have been informed, the permission was obtained from the administration of the hospital. The secrecy and confidentiality of the patients' record was assured for the officials of the hospital by promising to keep it secretly. In addition to keeping the data confidentially and secretly the data was not used for other purposes beyond the purpose of this research.

5. DISSEMINATION OF RESULTS

The findings of the study will be submitted to University of Gondar, institute of public health, Debirebirhan hospital, North shoa zone health department, Amhara regional health bureau and for those governmental organizations and non-governmental organizations interested in the subject matter. The finding will be presented in different conferences and workshops and will be sent to publication on scientific journal.

6. RESULTS

6.1. Characteristics of the registered patients

From the 10 years tuberculosis data of patients registered from October 2001 to June 2011 for treatment 4780 records were taken for analysis.

Table 2: The characteristics of patients registered for treatment in Debirebrhan hospital, October 2001 to June 2011.

Patient characteristics	Number (%)
Sex	
Male	2426(50.8)
Female	2354(49.2)
Residence	
Rural	2455(51.4)
Urban	2325(48.6)
Age categories	
0_14	672(14.1)
15_24	1160(24.3)
25_24	1357(28.4)
35_44	816(17.1)
45_54	460(9.62)
55_64	196(4.1)
65+	116(2.4)
Tuberculosis cases	
P/pos	1298(27.1)
P/neg	2847(59.6)
Extra pulmonary	635(13.3)
HIV test	
Test performed	1320(27.6)
Reactive	468(35.4)
Non-reactive	852(65.6)
Treatment category	

New	4373(91.5)
Relapse	97(2)
Failure	10(0.2)
Return after default	9(0.2)
Transferred in	152(3.2)
Others	139(2.9).

Figure 3 showed the distribution of patients in their age category. The colors on the bar chart indicate the share of each class in the specified age group. An exploratory analysis can also inform us the overall sense of the dataset like out layers and skewness.

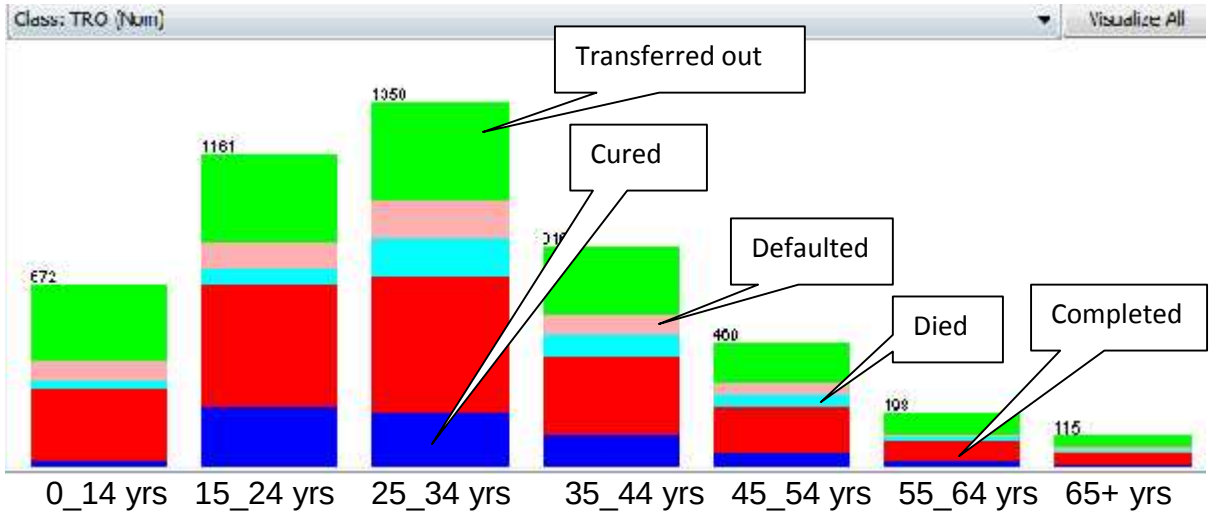

Figure 3- An exploratory analysis of age category of TB patients in Debirebirhan hospital

Table 3: Treatment outcome of patients registered for treatment in Debirebirhan hospital October 2001 to June 2011.

Treatment outcome	Number (%)
Cured	649(50)
Completed	1813(37.9)
Died	370(7.7)
Failed	4(0.3)
Defaulted	458(9.6)
Transferred out	1486(31.1).

6.2. Model building experiments

The data were 4780 patient records taken from the registration book of tuberculosis. After SMOTE the data 20 level the number of instances were reached to 25470. The decision tree model building by using 10 fold cross validation and percentage split with its default parameter the size of the tree and number of leaves became too long. So in order to limit the size of the tree and the number of leaves the parameter minimum number of instances (minNumObj) were adjusted from the default value by tried different alternative untill the the size of the tree and the number of the leaves reached to minimum. For MLP and SMO different parameters were tried to bring the better performance accuarcy for the model by changing the value of the parameter from the default values.

Experiment 1: Classification of records with J48 classifier using 10 fold cross validation test option

In order to classify the records based on their values for the attribute the model was trained using various parameters including the default parameters of the WEKA program until the performance of the model, the number of leaves and the size of the tree became relatively stable. A 10 fold cross validation was employed to partition the dataset into training set and test set; including the default values of the parameters.

When the model was developed using the default parameter with minimum number of instances (minNumobj) 2, number of folds (numfolds) 3 and reduced error pruning false 90.6% of instances classified correctly. But the number of leaves and the size of the tree were 508 and 888 respectively which were too long to understand the model and to generate rule. So as to reduce this complex and bushy tree the investigator attempted to modify the default values of the parameter. In order to achieved this objective the minimum number of objects changed to 80, number of folds to 5 and reduced error pruning to true which was 2, 3 and false in the default value respectively.

In reality, changing the default values has its own drawback as some of the instances are going to be classified incorrectly. In other words, records in a given leaf might be in different class and there could be attributes, which could further split the records in the same node into disjoint classes.

After modified some of the parameter as stated above the classifier output of J48 classifier was:

Number of Leaves:	61
Size of the tree:	94
Time taken to build model:	0.87 seconds

Evaluation of the developed model using J48 decision tree algorithm with 10 fold cross validation to test its performance

Correctly Classified Instances	21543	(84.6%)
Incorrectly Classified Instances	3927	(15.4%)
Kappa statistic	0.8118	
Mean absolute error	0.0717	
Root mean squared error	0.1905	
Relative absolute error	26.1%	
Root relative squared error	51.4%	
Coverage of cases (0.95 level)	99.1%	
Total Number of Instances	25470	

Figure 4 depicts while the J48 decision tree program was on training. As shown in the test options the method employed for partition data set into training and test was based on the cross-validation which was set by default. In addition to this, there are other options under Generic Object Editor to change the default values.

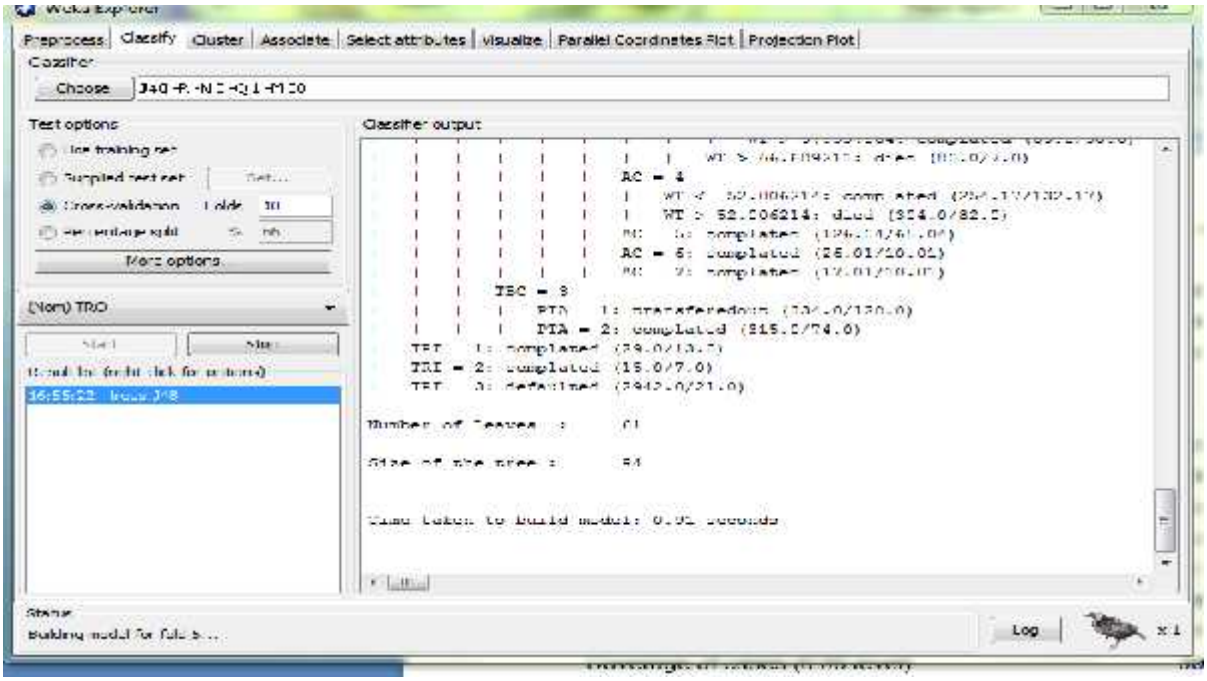

Figure 4- J48 decision tree algorithm on training

The detail accuracy of the classifier output illustrated that the values of different evaluation parameters in each class as shown in table 4.

Table 4- Detailed Accuracy by Class of J48 decision tree

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.998	0.002	0.993	0.998	0.996	0.998	Cured
0.563	0.05	0.649	0.563	0.603	0.922	Completed
0.876	0.096	0.734	0.876	0.799	0.951	Died
1	0	0.999	1	0.999	1	Failed
0.997	0.001	0.993	0.997	0.995	0.998	Defaulted
0.466	0.039	0.613	0.466	0.529	0.912	Transferred out
. 0.846	0.035	0.84	0.846	0.84	0.967	Weighted Avg.

As presented in table 5, the confusion matrix depicts that out of the total of 25470 records provided to the program, 21543 (85.6%) were classified correctly as a total accuracy of the model. The detail accuracy by each class is also presented as shown in the table. The result indicated that patients whose class is transferred out were classified with least accuracy (46.6%) as compared with other classes.

Table 5- Confusion Matrix of J48 decision tree

Actual	Predicted						Total	Accuracy in %
	cured	Completed	died	failed	defaulted	Transferred out		
Cured	5184	3	1	2	0	2	5192	99.86
Completed	38	2040	959	2	3	584	3626	56.26
Died	0	438	5188	0	12	282	5920	87.64
Failed	0	1	0	4095	0	0	4093	99.98
Defaulted	0	2	4	0	3652	6	3664	99.67
Transferred out	0	658	920	0	10	1384	2972	46.57
Total	5222	3142	7072	4099	3677	2258	25470	84.58

The investigator tried to improve the accuracy registered by the decision tree algorithms. Although different option is attempted by changing the values from the Generic Object Editor the result could not be improved more than 84.58%. The main challenge during this experimentation was the number of leaves and the size of the tree became increased when the performance improved. This is may be due to the characteristics of patients whose class was in transferred out similar with other classes especially with treatment completed and cured. As shown from the result also transferred out class carried out the highest incorrectly classified instances as compared to other classes.

Experiment 2: Classification of records with J48 classifier using percentage split test option of 65% for training and the remaining 35% for testing

The researcher tried different options to bring the better result by changing the default value of generic object editor. When the minimum number of objects and number of folds increased, the accuracy was increased but number of leaves and the size of the tree have been increased. Since increased number of leaves and size of the tree makes the decision tree none understandable and complex to generate the rule, experiment was done with changing the parameter value of minimum number of objects 80, number of folds 5 and reduced error pruning true to alleviate this problem.

The classifier output showed:

Number of Leaves:	61
Size of the tree:	94
Time taken to build model:	0.91 seconds

Evaluation of the developed model using J48 decision tree algorithm with percentage split of 65% training and 35% test.

Correctly Classified Instances	7404	(83.1%)
Incorrectly Classified Instances	1510	(16.9%)
Kappa statistic	0.7928	
Mean absolute error	0.0757	
Root mean squared error	0.1961	
Relative absolute error	27.6 %	
Root relative squared error	52.9 %	
Coverage of cases (0.95 level)	99.3%	
Total Number of Instances	8914	

Table 6 depicts experimental result of J48 decision tree algorithm with 65% train and 35% test option of percentage split.

Table 6- Detailed Accuracy by Class for J48 decision tree

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.997	0.002	0.993	0.997	0.995	0.998	Cured
0.462	0.041	0.641	0.462	0.537	0.91	Completed
0.861	0.116	0.691	0.861	0.767	0.945	Died
0.999	0	0.999	0.999	0.999	1	Failed
0.996	0.001	0.996	0.996	0.996	0.998	Defaulted
0.477	0.048	0.573	0.477	0.521	0.911	Transferred out
0.831	0.039	0.826	0.831	0.824	0.963	Weighted Avg.

Table 7 describes the confusion matrix of classifier output developed using J48 decision tree algorithm with percentage split. The total and detail accuracy is specified in relation to each class.

Table 7- Confusion Matrix for J48 decision tree with percentage split

Actual	Predicted						Total	Accuracy in %
	cured	Completed	died	failed	defaulted	Transferred out		
Cured	1865	1	1	2	0	1	1870	99.37
Completed	14	563	421	0	0	220	1218	46.22
Died	0	128	1778	0	3	156	2065	86.10
Failed	0	1	0	1439	0	0	1440	99.93
Defaulted	0	2	1	0	1250	2	1255	99.60
Transferred out	0	184	371	0	2	509	1066	47.75
Total	1879	879	2571	1441	1255	888	8914	83.06

Experiment 3: Classification of records with SMO (sequential minimal optimization) algorithm for support vector machine

SMO implements the sequential minimal optimization algorithm for training a support vector classifier, using polynomial or Gaussian kernels. Missing values are replaced globally, nominal attributes are transformed into binary ones, and attributes are normalized by default—note that the coefficients in the output are based on the normalized data. Normalization can be turned off, or the input can be standardized to zero mean and unit variance. Pair wise classification is used for multiclass problems. Logistic regression models can be fitted to the support vector machine output to obtain probability estimates. In the multiclass case the predicted probabilities will be coupled pair wise. When working with sparse instances, turn normalization off for faster operation.(6)

For this study case the default value were used to build the model.

The classification output showed:

Time taken to build model: 660.62 seconds

Evaluation of the developed model using SMO algorithm with cross validation test option

Correctly Classified Instances	20753	(81.5%)
Incorrectly Classified Instances	4717	(18.5%)
Kappa statistic	0.771	
Mean absolute error	0.2276	
Root mean squared error	0.3188	
Relative absolute error	82.9 %	
Root relative squared error	86 %	
Coverage of cases (0.95 level)	99.98 %	
Total Number of Instances	25470	

Experimental result of SMO algorithm with cross validation test option presented in table 8

Table 8- Detailed Accuracy by Class for SMO

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.999	0.002	0.993	0.999	0.996	0.998	Cured
0.322	0.012	0.811	0.322	0.461	0.731	Completed
0.979	0.203	0.594	0.979	0.739	0.893	Died
0.999	0	1	0.999	0.999	0.999	Failed
0.997	0.001	0.993	0.997	0.995	0.998	Defaulted
0.289	0.018	0.674	0.289	0.405	0.858	Transferred out
0.815	0.052	0.838	0.815	0.791	0.92	Weighted Avg.

In table 9 confusion matrix indicted patients whose class is transferred out classified with least accuracy as compared from other classes.

Table 9- Confusion Matrix for SMO

Actual	Predicted						Total	Accuracy in %
	cured	Completed	died	failed	defaulted	Transferred out		
Cured	5186	2	3	0	0	1	5192	99.88
Completed	32	1166	2059	2	3	364	3626	32.16
Died	0	62	5797	0	12	49	5920	97.92
Failed	3	1	0	4092	0	0	4096	99.90
Defaulted	0	5	5	0	3652	2	3664	99.67
Transferred out	1	201	1900	0	10	860	2972	28.94
Total	5222	1437	9764	4094	3677	1276	25470	81.48

Experiment 4: Classification of records with MLP (MultilayerPerceptron) algorithm for neural networks

MultilayerPerceptron is a neural network that trains using backpropagation. For the purpose of this study the researcher attempted to manage the output by changing different parameters from the genericobjecteditor.

Figure 6 shows the WEKA program is under running training with MLP algorithm. The neural network shown in the figure set with its default parameters by changing the GUI to true from the Generic Object Editor it was false by default.

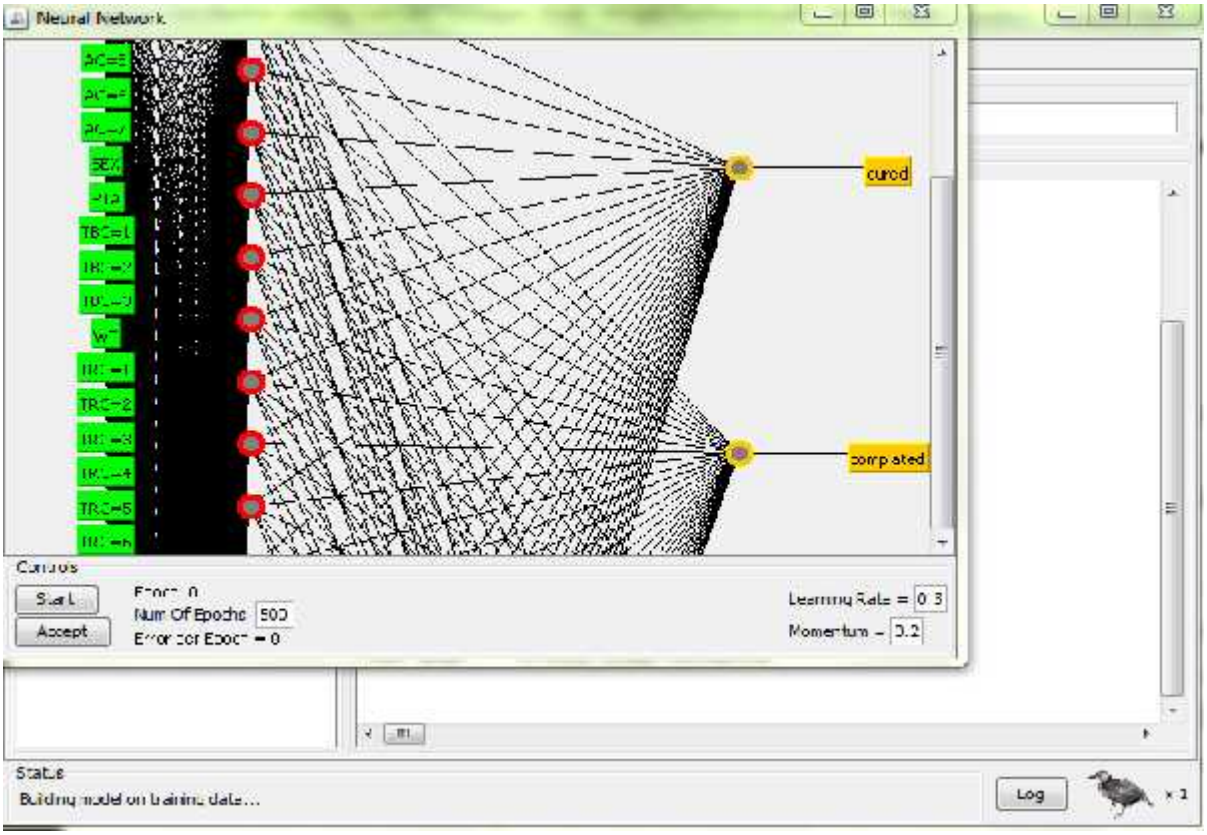

Figure 5- MLP algorithm under training

The classifier output showed:

Time taken to build model: 1705.98 seconds

Evaluation of the developed model using MLP algorithm with cross validation test option

Correctly Classified Instances	21854	(85.8%)
Incorrectly Classified Instances	3616	(14.2%)
Kappa statistic	0.8269	
Mean absolute error	0.0618	
Root mean squared error	0.1833	
Relative absolute error	22.5%	
Root relative squared error	49.5%	
Coverage of cases (0.95 level)	98%	
Total Number of Instances	25470	

The detail accuracy experiment result of MLP algorithm with cross validation test option depicted in table 8

Table 10-Detailed Accuracy by Class for MLP

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.998	0.001	0.994	0.998	0.996	0.999	Cured
0.622	0.05	0.673	0.622	0.646	0.932	Completed
0.88	0.08	0.769	0.88	0.821	0.962	Died
0.999	0	1	0.999	0.999	1	Failed
0.997	0.001	0.992	0.997	0.995	0.999	Defaulted
0.492	0.04	0.621	0.492	0.549	0.915	Transferred out
. 0.858	0.031	0.853	0.858	0.854	0.971	Weighted Avg.

In table 11 confusion matrix indicted patients whose class were transferred out classified with least accuracy as compared from other classes.

Table 11- Confusion Matrix for MLP

Actual	Predicted						Total	Accuracy in %
	Cured	Completed	died	failed	defaulted	Transferred out		
Cured	5181	7	1	1	0	2	5192	99.78
Completed	29	2254	804	0	4	535	3626	62.16
Died	0	344	5212	0	12	352	5920	88.04
Failed	0	3	0	4093	0	0	4096	99.93
Defaulted	0	4	4	0	3652	4	3664	99.67
Transferred out	0	737	760	1	12	1462	2972	49.19
Total	5210	3349	6781	4095	3680	2355	25470	85.80

6.3. Exploring pattern

The pattern of tuberculosis based on the attributes used in the study in relation to the treatment outcome as a class attribute can be explored from the rule generated from decision tree. Although MLP algorism can learn and understand the pattern of the data internally, it cannot visualize the rule as decision tree. Because of this shortcoming of MLP, the pattern was shown by using the rule generated in J48 decision tree with 10 fold cross validation which was registered in experiment one.

Generating Rules from Decision Tree

From the decision tree developed in the experiments, it is possible to find out a set of rules simply by traversing the tree and generating a rule for each leaf and making a combination of all the tests found on the path from the root to the leaf node. The generated rules can provide a decision support for the health professionals that working in tuberculosis control and treatment.

The following were some of the rules generated:

1. **If** sputum test result at 7th month=negative **then** cured (4182/36)
2. **If** sputum test result at 7th month=positive **then** failed (3279/3)
3. **If** sputum test result at 7th month=not applicable **and** treatment interruption >8 weeks **then** defaulted (2942/21)
4. **If** sputum test result at 7th month =not applicable **and** treatment interruption= 4 to 8 weeks **then** completed (15/7)
5. **If** sputum test result at 7th month =not applicable **and** treatment interruption <4 weeks **then** completed (29/13)
6. **If** sputum test result at 7th month =not applicable **and** treatment interruption= none **and** follow up sputum test at 7th month= no **then** completed (248/9)
7. **If** sputum test result at 7th month =not applicable **and** treatment interruption= none **and** follow up sputum test at 7th month= yes **then** transferred out (1)
8. **If** sputum test result at 7th month =not applicable **and** treatment interruption= none **and** follow up sputum test at 7th month= not applicable **and** tuberculosis case= pulmonary positive **and** HIV test offered= yes **then** transferred out (145/21)
9. **If** sputum test result at 7th month =not applicable **and** treatment interruption= none **and** follow up sputum test at 7th month= not applicable **and** tuberculosis case= extra pulmonary tuberculosis **and** patient address= rural **then** transferred out (334/128)
10. **If** sputum test result at 7th month =not applicable **and** treatment interruption= none **and** follow up sputum test at 7th month= not applicable **and** tuberculosis case= extra pulmonary tuberculosis **and** patient address= urban **then** completed (315/74).
11. **If** sputum test result at 7th month =not applicable **and** treatment interruption= none **and** follow up sputum test at 7th month= not applicable **and** tuberculosis case= pulmonary negative **and** HIV test result= reactive **and** ART started= no **then** completed (92/34)
12. **If** sputum test result at 7th month =not applicable **and** treatment interruption= none **and** follow up sputum test at 7th month= not applicable **and** tuberculosis

case= pulmonary negative **and** HIV test result= reactive **and** ART started= not applicable **then** died (54/4)

6.4. Comparison of experimental results

One of the purposes of this study was to compare the performance of different classification algorithms in classifying the tuberculosis patient's data set in Debirebirhan hospital and to identify the one which shows the better result. Accordingly, the experiments were carried out using decision tree with 10 fold cross validation and percentage split; SMO and MLP with 10 fold cross validation. The experiments of these studies indicated the performance accuracy of the MLP for artificial neural network was 85.8%, decision tree with 10 fold cross validation 84.6%, decision tree with percentage split 83.1% and SMO for support vector machine 81.5%. These results revealed that in the experiments all techniques have performed well even though the accuracy of the MLP exceeds. For practical application of the model, in terms of ease and simplicity to the user the decision tree is more self-explanatory and understandable. It generates rules that can be presented in simple human language. But the other two techniques are difficult to understand and apply with users understanding. These may need advanced users and software application to implement in decision making.

Table 12- Comparison of experimented results

Classifier	Accuracy	Kappa statistics	F-measure	ROC area	Incorrect classification
Decision tree with 10 fold cross validation	84.6%	0.8118	0.84	0.967	15.4%
Decision tree with percentage split	83.1 %	0.7928	0.824	0.963	16.9%
SMO (sequential minimal optimization)	81.5%	0.771	0.791	0.92	18.5%
MLP (multilayerperceptron)	85.8%)	0.8269	0.854	0.971	14.2%

.

7. DISCUSSION

7.1. The model

The aim of this study is to identify patterns from the tuberculosis data, to develop the model that can predict the future pattern from the new data and to identify an appropriate data mining algorithm.

In this study there was 4 experiments undertaken to develop the predictive model, using MLP, decision tree with 10 fold cross validation, decision tree with percentage split and SMO the accuracy of these models were 85.8 %, 84.6 %, 83.1 %, and 81.5% respectively.

A study conducted in India on performance analysis of data mining techniques for placement chance Prediction using 300 records showed the accuracy for Multilayer perceptron, J48 decision tree and SMO were 87.4%, 84.9% and 84% respectively(18). Multilayer perceptron was showed the best performance when it compared with others. The performances of these algorithms were similar to this study in relation to its level of performance. But MLP performed higher than found in this study. The difference of accuracy might be due to this study used larger dataset which were 4780 records.

Another study done in India to classify the tuberculosis patient's pattern using 700 records with classifier algorithms of support vector machine, artificial neural network and decision tree the performance accuracy was 98.7%, 93% and 92.4% respectively. The accuracy found from the study showed a discrepancy from this study. This was may be due to the difference of number of instances, number of attributes and type of class attributes used. The values for number of instances, number of attributes and type of class attributes were 4780, 20 and multivariate in this study and 700, 12 and bivariate in the sited study respectively.

From the experiments done in this research using algorithms of decision tree with 10 fold cross validation and percentage split, MLP and SMO, multilayer perceptron (MLP) showed the best performance when compared to others. Even if decision tree algorithm is the better one on its understandability and ease of application to the domain expert without the need of highly qualified data mining expert and the help of

software, MLP prevailed over decision tree for the classification of tuberculosis dataset used in this study.

To build the neural network model, WEKA software package is used and it employs back propagation algorithm in developing a model. All the 19 attributes which was feed in to the input node of the neural network after normalized in to -1 and 1 automatically found predictor. Each input attributes are passed through the hidden node and output node with weighted lines. The outputs again back propagate as an input through the network by calculating the error until the error reached to minimum. As well multilayer perceptron used in this research classified correctly 21854 (85.8%) instances and classified incorrectly 3616 (14.2%) instances.

7.2. Exploring the pattern

Sine MLP cannot visualize the rule explicitly with understandable human language, the rule developed using J48 decision tree in experiment one was used for the purpose of showing the pattern. From the developed decision tree model it was seen that the attributes like sputum test result at 7th month, treatment interruption, follow up sputum test at 7th month, tuberculosis case, HIV test result, patient address and ART started were the most determinant predictor attributes. This can be proved from the decision tree algorithm that brought these attributes at the top of the developed tree.

The important finding discovered from this study was performing follow up sputum test at 7th month for pulmonary positive cases is the key predictor attribute to identify patients who were cured and treatment failed. Because the attribute sputum test results at 7th month (STR7M) become a root node in the developed decision tree. Tuberculosis now days become a critical problem because of its multidrug resistance. So in order to diagnose whether there is multidrug resistance or not conducting follow up sputum test is fundamental. That means if the previously smear positive patient tested sputum positive at 7th month of treatment, s/he declared as treatment failed. This is important to decide for retreat the patient with treatment category II or to send for appropriate microbiological testing to check multidrug resistance. This shows that to declare the patient as cured or treatment failed,

performing follow up sputum test is crucial. If the health professionals didn't know at least 2 follow up sputum test result of the patient including the 7th month, they may face a problem to decide whether the patient is cured or treatment failed. For instance if follow up sputum test result of an initially pulmonary positive patient at 7th month (mandatory test) not done, the patient can be declared as completed but that cannot guarantee his/ her cured.

All the attributes used in this study were considered as a predictor attributes to explore the pattern and predict treatment outcome based on its information gain. This was tested using WEKA explorer attribute selection option by choosing information gain attribute evaluation.

7.3. Selection and evaluation of the model

From four experiments conducted using the specified algorithms Multilayer perceptron showed the highest accuracy when compared to with others. So the model developed using this algorithm was the best for the classification and prediction of tuberculosis data. The developed model with multilayer perceptron was evaluated by different evaluation parameters of the software and showed kappa statistic of 0.826, recall of 0.858, precision of 0.853 and ROC area of0.971.

This shown us the developed model has an agreement of 82.6% between actual and predicted, true positive rate of 85.8% and predict instances as positive with probability of 0.971.

8. STRENGTH AND LIMITATION OF THE STUDY

8.1. Strength of the study

This research attempted to explore the pattern of tuberculosis in relation to its treatment outcome. The models developed in this research can predict the future pattern of TB treatment outcome. The outputs found from this research can provide the decision making support for the health professionals, public health officials and decision makers. The scope of this research is to illustrate the potential applicability of data mining for health care activities to make easy decision making and to identify the potential risk area of public health. This research can be applied in tuberculosis prevention and control aspects in Debirebirhane hospital and other similar areas.

8.2. Limitations of the study

Even if data mining is done in most cases from data bases, due to unavailability of data base for tuberculosis registration in Debirebirhan hospital the researcher has been done this research by converting the hard copy to excel file. This task was consumed much of the research time. In its nature conducting data mining research takes 60% of the research time for data preparation (7). Since the data for this study was taken from registration books some important attributes like sign and symptoms of the patients couldn't found. There was a big challenge getting related research or literature done on this topic.

9. CONCLUSION AND RECOMMENDATION

9.1. Conclusion

.All algorithms experimented in this study showed a promising result to classify tuberculosis data and to develop predictive model. This study revealed that the pattern of Tuberculosis can be successfully explored as cured, completed, died, failed defaulted and transferred out in relation to its treatment outcome. The result also indicated data mining technique can be applied to develop models that can predict the pattern of tuberculosis. From the developed models Multilayer perceptron algorithm showed the best accuracy when compared to decision tree and sequential minimal optimization algorithms. Sputum test result of 7^{th} month for smear positive patients was the most determinant predictor attribute for cured and failed classes. All attributes feed for the algorithms were considered as significant attributes. The cure and completion rate of tuberculosis treatment (treatment success rate) accounted 51.5% which was unsatisfactory. The undesirable result of tuberculosis treatment died, treatment defaulted and failed accounted 17.4% which is serious problem as a public health concern. This study also showed data mining is an appropriate tool to describe the distribution of treatment outcome in each class. To sum-up, data mining is an appropriate technique to apply for the classification of tuberculosis patients using the historical information to predict the future pattern.

9.2. Recommendation

The results of this study were found to be promising to address practical problems in prevention and control of tuberculosis infection. In order to make evidence based decision making on tuberculosis prevention and control applying data mining technique might play a great role.

According to the pattern and other findings from this research the following recommendations are forwarded.

- ✓ The hospital need to have electronic records or database for tuberculosis registration to assure the quality of data and easy access of information.
- ✓ To alleviate the problem associated with treatment failure the health professionals need to give attention on sputum follow up test at least 2 of the test including the 7^{th} month for pulmonary positive cases to declare as cured or treatment failed.
- ✓ The health care providers and other concerned governmental and nongovernmental sectors need to work on reducing the number of defaulters by making awareness to the community.
- ✓ Since HIV test result is one of the determinant attribute health professionals need to strengthen screening tuberculosis patients for HIV in order to take appropriate measure to save the life of the patient.
- ✓ The health care providers and officials of the hospital need to know the determinant attributes that affect treatment outcome of tuberculosis to identify the focus area during treatment, to design effective resource allocation and training for the prevention and control of tuberculosis.
- ✓ Further research will also be expected to be undertaken on large scale data and adding attributes like sign and symptom of the patients.

10. REFERENCES

1. World health organization. Treatment of tuberculosis: guidelines. fourth ed2009. 160 p.
2. World health organization. Global tuberculosis control 2011. Global Report. France: WHO, Press W; *2011* 12/16/2011 Report No.: 16.
3. Maniya H, Hasan MI, Patel KP. Comparative study of Naïve Bayes Classifier and KNN for Tuberculosis. International Conference on Web Services Computing (ICWSC); India. India: International Journal of Computer Applications (IJCA); 2011. p. 1-5.
4. Federal Ministry of Health Ethiopia. Health Sector Development Program IV 2010/11 – 2014/15. Addis Ababa, Ethiopia: FMOH; 2010. p. 11.
5. Sumathi S, Sivanandam SN. Introduction to Data Mining and its Applications. Berlin: Springer; 2006 [cited 2011 12/17/2011]. Available from: www.springer.com.
6. Witten Ih, Frank E. Data mining Practical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann Publishers; 2005 [cited 2012 2.21/2012]. Available from: www.books.elsevier.com.
7. Pyle D. Data Preparation for Data Mining [4/20/2012]. USA, Sanfransisko: Morgan Kaufmann Publishers, Inc.; 1999 [cited 2012 5/21/2012]. Available from: http://www.mkp.com.
8. Two Crows Corporation. Introduction to Data Mining and Knowledge Discovery. U.S.A: Two Crows Corporation; 1999 [cited 2011 12/17/2011]. Available from: www.twocrows.com.
9. Ranjan J. Application of Data mining Technique in the Field of Pharmaceutical Industry. India: Journal of Theoretical and Applied Information Technology; 2007 [cited 2012 2/21/2012]; Available from: www.jatit.org.
10. Asha T, Natarajan S, Murthy K. A Hybrid Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classification. Journal of Computimg Press. APRIL 2011;3(4):185-92.
11. Canlas RD. Data Mining in healthcare:current application and issue [Master Thesis]. Australia: Carnegie Mellon University; 5 August 2009.

12. Freitas AA, Vasieva O, Magalhães JPd. A data mining approach for classifying DNA repair genes into ageing-related or non-ageingrelated. BMC Genomics. 2011;12:27.

13. Yumuşak N. Tuberculosis Disease Diagnosis Using Artificial Neural Network Trained with Genetic Algorithm. J Med Syst. 2009;35:329–32.

14. Periwal V, Rajappan JK. Predictive models for anti-tubercular molecules using machine learning on high-throughput biological screening datasets. BioMed center. 18 November 2011;4:504.

15. Djam XY, Kimbi YH. A Decision Support System for Tuberculosis Diagnosis. The Pacific Journal of Science and Technology. November 2011;12(2):410-25.

16. Tessema B, Muche A, Bekele A, Reissig D, Emmrich F, Sack U. Treatment outcome of tuberculosis patients at Gondar University Teaching Hospital, Northwest Ethiopia. A five - year retrospective study. BioMed center Public Health. 2009;9:371.

17. Han H, Wang W-Y, Mao B-H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Springer. 2005:878 – 87.

18. V.Ramesh PP, P.Yasodha. Performance Analysis of Data Mining Techniques for Placement Chance Prediction. August-2011;2(8):7.

19. Federal Ministry of Health Ethiopia. Tubercilosis, leprosy and TB/HIV prevention and control program In: Diseases prevention and control FMoH, editor. Fourth ed. Addis Ababa< Ethiopia: FMOH; 2008. p. 218.

Annex 1_ Description of the dataset

Cured _ An initially smear positive patient who is sputum smear negative at, or one' month' prior to, the completion of treatment and on at least one previous occasion (usually at the end of the 2nd and 5th month).

Treatment completed_ A patient who completed treatment but for whom smear results are not available at 7th or one month prior to completion of treatment.

Died _ A person who dies for any reason during the course of treatment.

Treatment failed_ A patient who remains or become again smear positive at the end of 5th 'month' or late during treatment, Or a patient who was P/NEG at the beginning and returned out smear positive at the end of the 2nd month.

Defaulter_ A patient who has been on treatment at least for four weeks and whose treatment was interrupted for 8 or more consecutive weeks.

Transfer out_ A patient who started treatment and has been transferred out to another reporting unit and for whom the treatment outcome is not known at the time of evaluation of treatment results.

Treatment success_ the sum of patients who are declared "cured" and "treatment completed"(19).

Diagnosis category by type of patients

Category I: consists mainly of new, smear-positive tuberculosis cases, but includes new smear-negative cases with extensive parenchymal lesions, and new cases with severe extra pulmonary tuberculosis (disseminated, meningeal, pericardial, peritoneal, bilateral pleural, spinal, intestinal and genito-urinary). A new case is defined as a patient who has never previously been treated for tuberculosis or who has received treatment for less than one month.

Category II: smear-positive cases who have already received treatment for at least one month in the past and who need to receive re-treatment. Among these patients

three groups can be distinguished in: “Relapses” – patients who have been treated and declared cured, but whose smear examinations are once again positive.
“Failures” – patients whose smear examinations have remained positive or have once again become positive five or more months after starting treatment. “Return after interruption” – patients who return to the health centre smear-positive after interrupting treatment for more than two consecutive months.
Category III: new cases of smear-negative pulmonary or extra pulmonary tuberculosis (excluding those with severe forms, included in Category I) who have never previously been treated for as much as one month in the past.
Category IV: chronic cases defined as smear-positive cases of pulmonary tuberculosis who have previously received a supervised re-treatment regimen.

Definition of Tuberculosis cases

Smear positive pulmonary TB (P/POS TB)_ A patient with at least two initial sputum smear examinations positive for AFB by direct microscopy, OR a patient with one initial smear examination positive for AFB by direct microscopy and culture positive, OR a patient with one initial smear examination positive for AFB by direct microscopy and radiographic abnormalities consistent with active TB as determined by a clinician(19).

Smear negative pulmonary TB (P/NEG TB)_ A patient having symptoms suggestive of TB with at least 3 initial smear examinations negative for AFB by direct microscopy, and

- No response to a course of broad spectrum antibiotics, and
- Again 3 negative smear examinations by direct microscopy, and
- Radiological abnormalities consistent with pulmonary tuberculosis, and
- Decision by a clinician to treat with a full course of anti-tuberculosis.
  Or a patient his diagnosis is based on culture positive for M. tuberculosis but 3 initial smear examinations negative by direct microscopy(19)

Extra pulmonary TB (EP TB)_ TB in organs other than lungs, proven by one culture positive specimen from an extra pulmonary site or histo-pathological evidence from a biopsy, OR TB based on strong clinical evidence consistent with active EPTB and the decision by the physician to treat with a full course of ant-TB therapy(19).

Definition of treatment category

New (N)_ A patient who never had treatment for TB, or has been on previous anti-TB treatment for less than four weeks.

Relapse (RL)_ A patient declared cured or treatment completed of any form of TB in the past, but who reports back to the health service is now found to be AFB smear positive or culture positive.

Treatment failure (TF)_ A patient who, while on treatment, is smear positive at the end of fifth month or late, after commencing. Treatment failure also includes a patient who was sputum smear initially negative but who becomes smear positive during treatment.

Return after default (RD)_ A patient previously recorded as defaulted from treatment and return to the health facility with smear positive sputum .

Transfer in (TI)_ A patient who started treatment in one treatment unit and is transferred in to another treatment unit to continue treatment.

Others (O)_ A patient who does not fit any of the above mentioned categories (e,g. a P/NEG TB patient who return after treatment interruption)(19).

Definition of HIV test result

Reactive (R)_ a TB patient who tested for HIV and became positive or an HIV positive patient registered for TB treatment.

Non-reactive (NR)_ a TB patient who tested for HIV and became negative.

Indeterminate (I)_ a TB patient who tested for HIV but the result not exactly known

Annex 2_Information Sheet and Consent Form

Title of the Research Project

Application of Data Mining to Explore the Pattern of Tuberculosis: The Case of Debire Birhan Hospital, North Shoa, Ethiopia

Name of Principal Investigator: Mengistu Yilma

Name of the Organization: Institute of Public Health, College of Medicine and Health Sciences, University of Gondar

Name of the Sponsor: Amhara National Regional state Health Bureau.

Information Sheet and consent form prepared for application of data mining to explore the pattern of tuberculosis.

Introduction

This information sheet and consent form is prepared in order to explaining the aim of the research project. The main aim of the research project is to explore patterns from the tuberculosis data and develop predictive model using data mining technology. The research group includes 6 trained IT diploma, 2 trained BSC health professionals and two advisors one from University of Gondar and one from Addis Ababa university,.

Purpose of the Research Project

The intent of the research project is to explore patterns from the tuberculosis data and develop predictive model using data mining technology. Data mining technology can be applied to evaluate the effectiveness and the outcomes of patient groups treated with different drug regimens for the same disease or condition can be compared to determine which treatments work best and are most cost-effective(11). It is therefore the aim of this study to explore important pattern from tuberculosis data, build and test the predictive model by using this historical data.

The researcher initiated to conduct this study

- Since there was no study conducted previously on application of data miming to explore the pattern of tuberculosis in Debirebirhan hospital.
- Data mining is powerful, applicable and feasible tool to automate and up-to-date the healthcares decisions

Procedure

The whole records starting from the beginning of DOTS (Directly Observed Treatment Short course) of tuberculosis patients registered for TB treatment in Debirebirhan hospital from October, 2001 to June, 2011 are the principal target dataset for this study.

After the purpose and objective of the study have been informed, the permission will be obtained from the administration of the hospital. The secrecy and confidentiality of the patients' record will be assured for the officials of the hospital by promising to keep it secretly. In addition to keeping the data confidentially and secretly the data will not be used for other purposes beyond the purpose of this research.

Risk and /or Discomfort

Since there are no individuals participating physically for this research not the time and discomfort issue raised.

Benefits

By using the historical data of the hospital it is possible to explore the pattern of the tuberculosis patient and developing the predictive model that can predict the pattern of future patients. This helps the health professional to support their decision for easy diagnosis and treatment.

Confidentiality

The record of tuberculosis patients accessed for the purpose of this research will be kept secretly and no information can be exposed for anyone except the research group.

Person to contact

This research project will be reviewed and approved by the ethical committee of the University of Gondar. If you want to know more information you can contact the committee through the address below. If you have any question you can contact any of the following individuals and you may ask at any time you want.

1. Mr. Mengistu Yilma

Mobile: +251913593241 / e-mail newmany55@gmail.com

2. Mr. Takele Tadesse University of Gondar

Mobile: +251918773317 /e-mail takele_tadesse@yahoo.com

3. Dr. Million Meshesha Addis Ababa University

e-mail meshe84@gmail.com

የመረጃና የስምምነት ውል ቅፅ

የምርምሩ/ጥናቱ ርዕስ

በኢትዮጵያ፡ ሰሜን ሸዋ ዞን መስተዳድር በደብረ ብርሃን ሆስፒታል የቲቢ በሽታ ስርጭትን ለማወቅና የወደፊቱን ስርጭት ለመተነበይ ከብዙ መረጃ ውስጥ ጠቃሚ እውቀት ማውጣት የሚያስችል ዘዴ በመጠቀም ምርምር ማካሄድ፡፡

የዋና ተመራማሪው ስም፡ መንግስቱ ይልማ

የድርጅቱ ስም: በጎንደር ዩኒቨርሲቲ ህክምናና ጤና ሳይንስ ኮላጅ የህብረተሰብ ጤና አጠባበቅ ት/ቤት

ወጪውን የሚሸፍነው ፡ **የአማራ ብሔራዊ ክልሊዊ መንግስት ጤና ጥበቃ ቢሮ**

የቲቢ በሽታ ስርጭትን ለማወቅና የወደፊቱን ስርጭት ለመተንበይ ከብዙ መረጃ ውስጥ ጠቃሚ እውቀት ማውጣት የሚያስችል ዘዴ በመጠቀም ምርምር ለማካሄድ የተዘጋጀ የመረጃና የስምምነተ ውል ቅፅ፡፡

መግቢያ

ይህ የመረጃና የስምምነት ውሌ ቅፅ የተዘጋጀው የምርምሩን ዋና አላማ ለማሳወቅ ነው፡፡ የዚህ ምርምር ዋና አላማ የቲቢ በሽታ ስርጭትን ለማወቅና የወደፊቱን ስርጭት ለመተንበይ ከብዙ መረጃ ውስጥ ጠቃሚ እውቀት ማውጣት የሚያስችል ዘዴ በመጠቀም ምርምር ማካሄድ ሲሆን የዚህ ምርምር ስራ ቡድን አባል የሚሆኑት 6 በመረጃ ቴክኖሎጂ ዲፕሎማ ያላቸው፡ 2 በጤና ሳይንስ ዲግሪ ያላቸው ባለሙያዎች እና 2 የምርምሩ አማካሪውች ናቸው፡፡

የጥናት ፕሮጀክቱ የሚካሄዴበት ምክንያት

የጥናቱ ዓላማ የቲቢ በሽታ ስርጭትን ለማወቅና የወደፊቱን ስርጭት ለመተነበይ ከብዙ መረጃ ውስጥ ጠቃሚ እውቀት ማውጣት የሚያስችል ዘዴ በመጠቀም ምርምር በማካሄድ የጤና ባለሙያዎች ህመምተኞችን ለመመርመርና ለማከም በሚያደርጉት ሂድት አጋዥ ሁኔታ ለምፍጠር ነው፡፡ በሌላ በኩል የዚህ አይነት ምርምር በደብረ ብረሃን ሆስፒታል ተካሂዶ አለማወቁ እና ምረምሩ ጠቃሚና ሊተገበር የሚችል በመሆኑ ለማካሄድ ታስቧል፡፡

የጥናቱ አካሂር

ይህ ጥናት የሚካሄደው ከጥቅመት 1993 እስከ ሰኔ 2003 ለቲቢ ምርመራ ከተመዘገቡ ህሙማን መዝገብ ላይ ነው፡፡ ጠናቱ በሚሰጠው ጠቀሜታ ዙሪያ ገለፃ በማድረግ

ከሆስፒታሉ ሃላፊውች ፍቃድ በማገኘት ይካሄዳል። የመረጃው ደህንነትና ሚስጥራዊነት ለመጠበቅ ቃል በመግባት ይፈፀማል። ከጥናቱ ስራ ውጭ ለሌላ አገልገሎትም አይውልም።

ሊደርስ የሚችል ጉዳትና አለመመቸት

በጥናቱ በቀጥታ የሚሳተፍ ሰው ባለመኖሩ የጊዜና ያለመመቸት ጥያቄዎች አይነሱም።

ጠቀሜታ

የሆስፒታሉን የቆየ መረጃ በመጠቀም የቲቢ በሽታ ስርጭትን ለማወቅና የወደፊቱን ስርጭት በመተንበይ የጤና ባለሙያዎች ህመምተኞችን ለመመርመርና ለማከም በሚያደርጉት ሂድት አጋዥ ሁኔታ ለምፍጠር ያስችላል።

ምስጢራዊነት

ለዚህ ጥናት ሲባል የሚገኘው መረጃ በጥነቱ ስራ ውስጥ ከሚሳተፉት ቡኖች በስተቀር ማንም በማያገኘው ሁኔታ በምስጢር ይቀመጣል።

ሊያገኙዋቸው የሚችሉ ሰዎች

ይህ የምርምር ፕሮጀክት በጎንደር ዩኒቨርሲቲ የስነ ምግባር ኮሚቴ ተከልሶ የሚፀድቅ ይሆናል። የበለጠ መረጃ ማግኘት የሚፈልጉ ከሆነ ኮሚቴውን በሚከተሉት አድራሻዎች ማግኘት ይችላሉ። የትኛውም ዓይነት ጥያቄ ሲኖርዎት ከዚህ ቀጥሎ የተጠቀሱትን ግለሰቦች ማግኘትና በማንኛውም ጊዜ መጠይቅ ይችላሉ።

1. አቶ መንግስቱ ይለማ
   የእጅ ስለክ 0913593241/ E-mail newmany55@gmail.com
2. አቶ ታከለ ታደሰ፡- ጎንደር ዩኒቨርሲቲ

   የእጅ ስለክ 0918773317/ E-mail takele_tadesse@yahoo.com

3. ዶ/ር ሚልዮን መሸሻ፡- አዲስ አበባ ዩኒቨረሲቲ
   E-mail meshe84@gmail.com

Annex 3- Sample J48 pruned tree

STR7M = 0: cured (4182.0/36.0)

STR7M = 1: failed (3279.0/3.0)

STR7M = 2

| TRI = 0

| | FST7M = 0: completed (248.0/9.0)

| | FST7M = 1: transferedout (1.0)

| | FST7M = 2

| | | TBC = 1

| | | | HIVTO = 0

| | | | | PTA = 1

| | | | | | AC = 1: transferedout (9.0)

| | | | | | AC = 2: transferedout (126.0/42.0)

| | | | | | AC = 3: died (262.0/81.0)

| | | | | | AC = 4: died (78.0/29.0)

| | | | | | AC = 5: died (53.0/15.0)

| | | | | | AC = 6: transferedout (6.0)

| | | | | | AC = 7: transferedout (2.0)

| | | | | PTA = 2: died (643.0/67.0)

| | | | HIVTO = 1: transferedout (145.0/21.0)

| | | TBC = 2

| | | | HIVTR = 0

| | | | | PTA = 1: transferedout (513.0/187.0)

| | | | | PTA = 2: completed (275.0/45.0)

| | | | HIVTR = 1

| | | | | ART = 0: completed (92.0/34.0)

| | | | | ART = 1

| | | | | | WT <= 49.557305

| | | | | | | WT <= 39.019804: transferedout (84.0/27.0)

| | | | | | | WT > 39.019804: died (555.0/149.0)

| | | | | | WT > 49.557305: transferedout (94.0/39.0)

| | | | | ART = 2: died (54.0/4.0)

| | | | HIVTR = 2

Declaration

I, the undersigned, senior MPH student declare that this thesis is my original work in partial fulfillment of the requirement for the degree of Master of Public Health.

Name: Mengistu Yilma

Signature: _____

Place of submission: Institute of public Health, College of Medicine and Health Sciences, University of Gondar

Date of Submission: _____

This thesis work has been submitted for examination with my/ our approval as university advisor(s).

Advisors

Name	Signature	**Date**
1. Mr. Takele Tadesse (BSC, MPH)	_____	_____
2. Dr. Million Meshesha (PHD)	_____	_____